Descriptive Statistics

GENOME 560, Spring 2026

Saori Sakaue (sakaue@uw.edu)

What Are Descriptive Statistics?

Basic numerical summaries of data

Why Do We Collect Data?

- To **describe** characteristics of a sample or samples
- To make **inferences** about a population

Central Dogma of Statistics

$$\text{Population} \xrightarrow{\text{sampling}} \text{Sample} \xrightarrow{\text{inference}} \text{Population}$$

We use **samples** to make **inferences** about the **population**.

Why Descriptive/Graphical Summary?

Before making inferences from data, it is essential to understand its basic structure.

We look to the data:

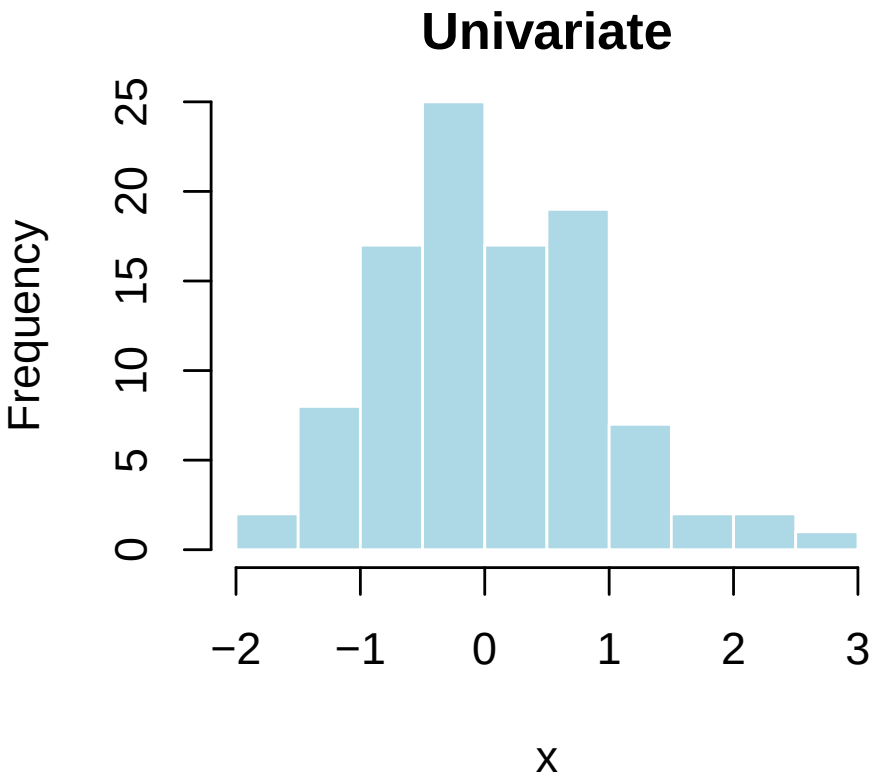
- To catch mistakes
- To see patterns
- To find violations of statistical assumptions
- To generate hypotheses

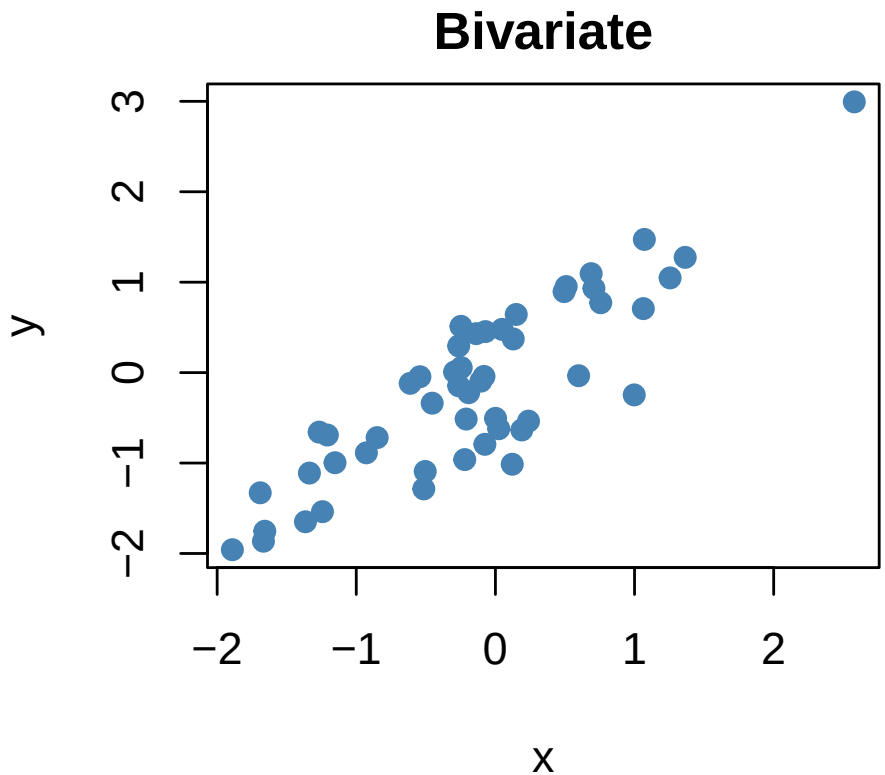
Types of Data

Quantitative (Numerical)	**Qualitative (Categorical)**
Continuous (e.g., height, weight)	Nominal (e.g., blood type)
Discrete (e.g., counts)	Ordinal (e.g., disease stage)

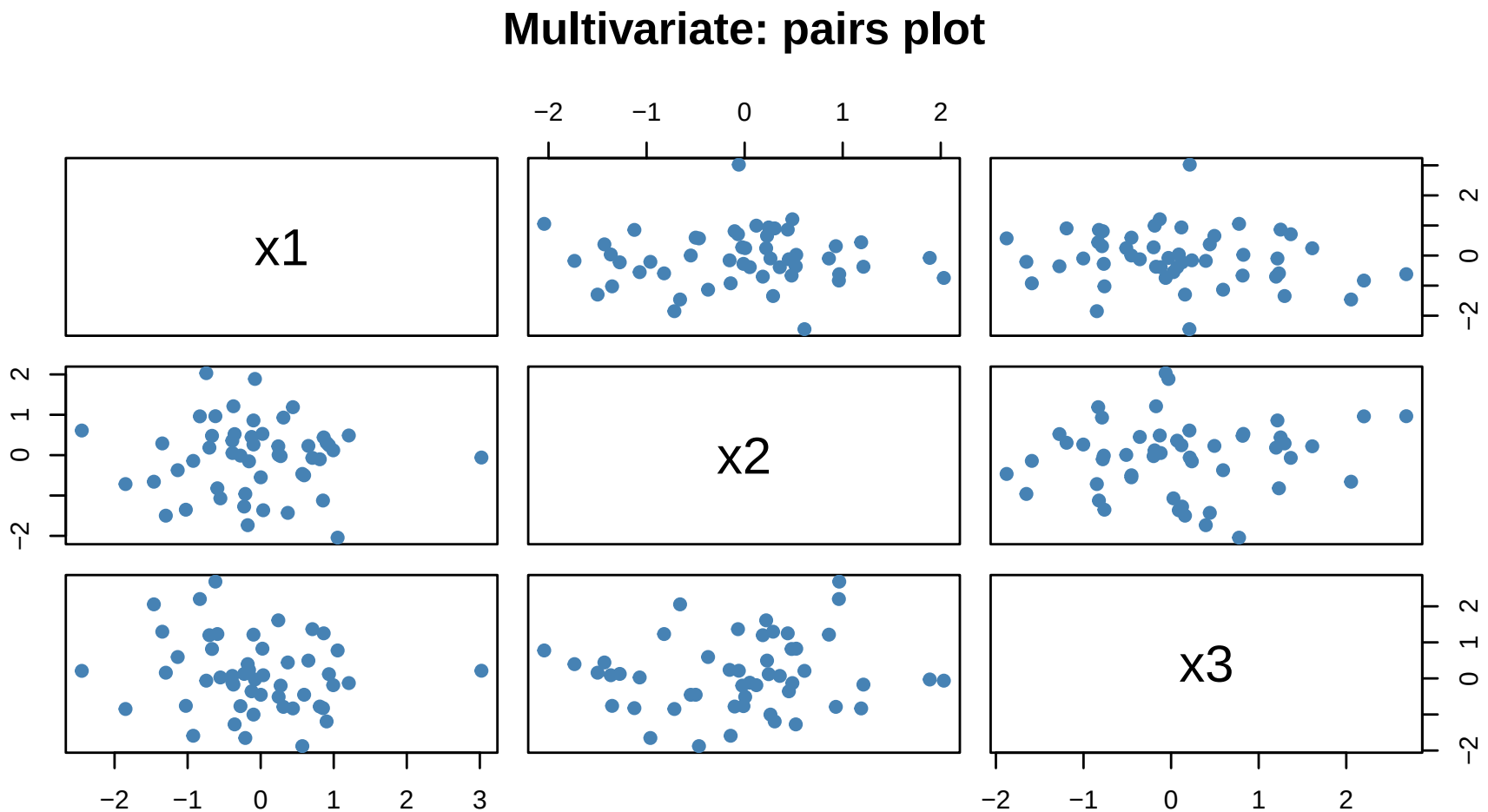
Dimensionality of Data Sets

- **Univariate:** one variable per subject
- **Bivariate:** two variables per subject
- **Multivariate:** many variables per subject





Dimensionality: Multivariate



Random Variables (RV)

- **Population:** the entire group of interest
- **Sample:** a subset of the population

A **random variable** is a numerical outcome of a random process.

Random Variables (RV)

- **Discrete RV:** countable number of values (e.g., number of mutations)
- **Continuous RV:** any value in an interval (e.g., gene expression level)

Section 1

Central Tendency

Numerical Summaries of Data

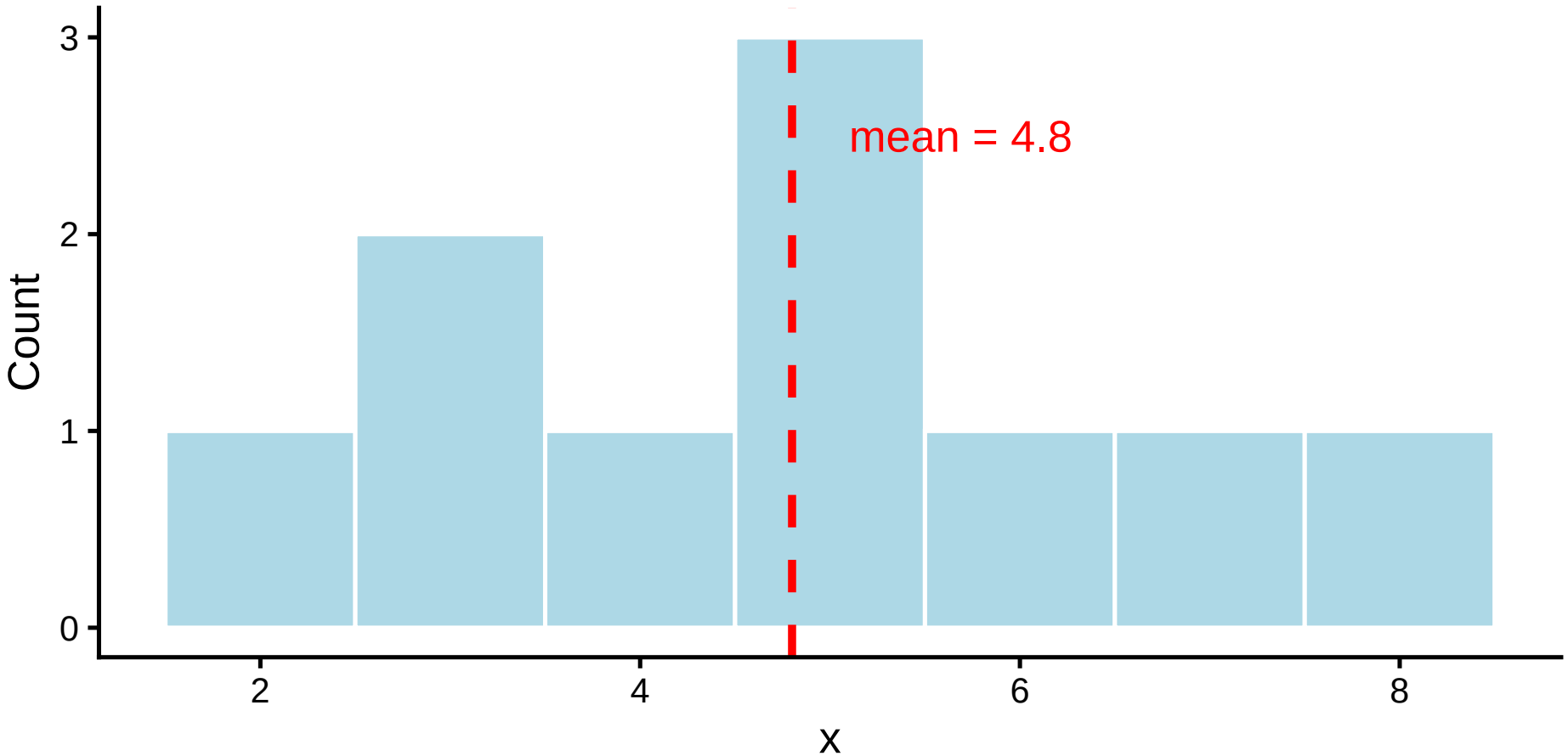
- ***Central Tendency measures.*** They are computed to give a "center" around which the measurements in the data are distributed.
- ***Variation or Variability measures.*** They describe "data spread" or how far away the measurements are from the center.
- ***Relative Standing measures.*** They describe the relative position of specific measurements in the data.

Central Tendency Measures: Mean

To calculate the mean, add values and divide by the number of observations:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Central Tendency Measures: Mean (continued)



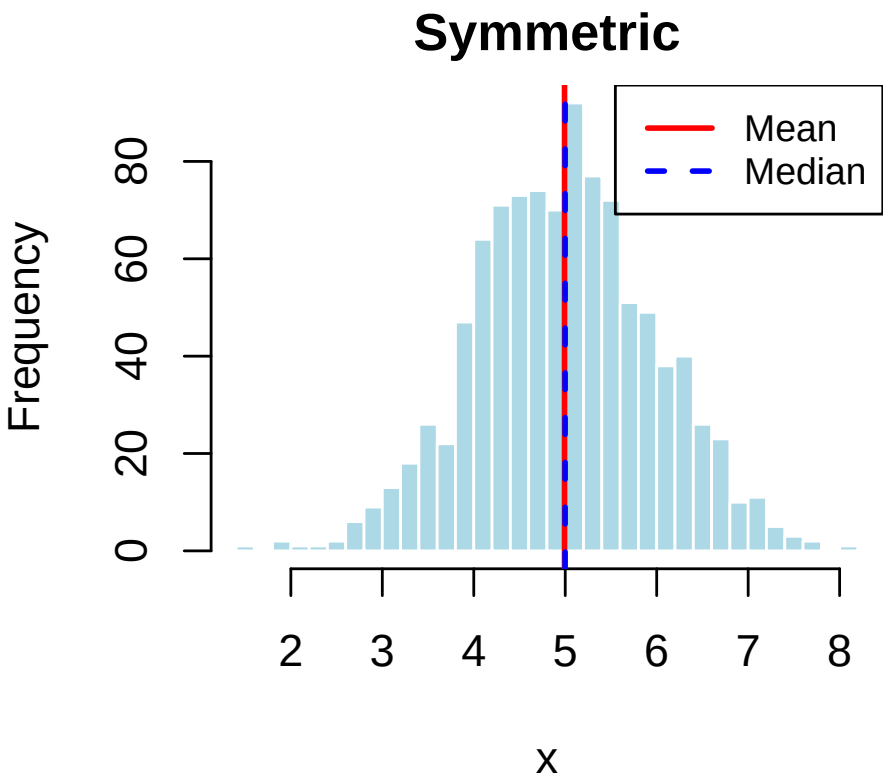
Central Tendency Measures: Median

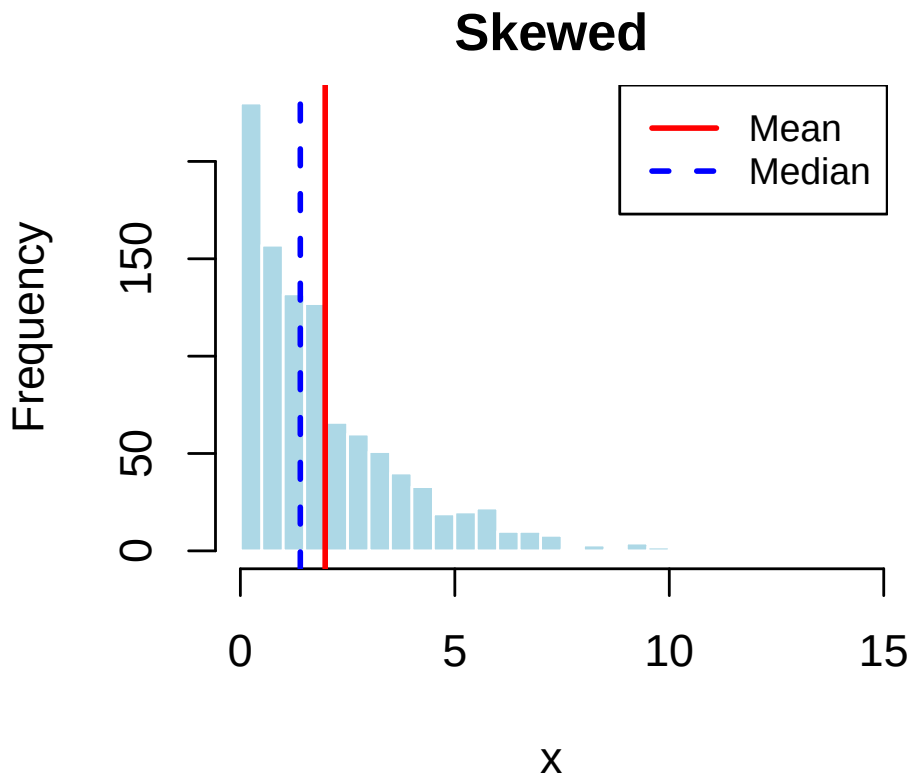
The **median** is the middle value when data are ordered.

- For odd n: the middle observation
- For even n: average of the two middle observations

When To Use Which?

- **Mean** is useful for symmetric distributions without outliers
- **Median** is useful for skewed distributions or data with outliers





Section 2

Variability

Variability Measures: Variance

Average of squared deviation of values from the mean.

Population Variance (if you can observe the entire space):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Variability Measures: Variance

Derivation: $\sigma^2 = E[(X-\mu)^2] = E[X^2] - (E[X])^2$

$$\sigma^2 = \underbrace{E[X^2]}_{\text{mean of squares}} - \underbrace{(E[X])^2}_{\text{squared mean}}$$

Variability Measures: Sample Variance

Unbiased estimator of variance:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Why Dividing by $n-1$?

$\bar{x}$ is estimated from the sample, not the true population mean μ.

$$E\left[\frac{1}{n}\sum(x_i-\bar{x})^2\right]=\frac{n-1}{n}\sigma^2\neq\sigma^2$$

Dividing by $n-1$ (**Bessel's correction**) gives an unbiased estimator:

$$E\left[\frac{1}{n-1}\sum(x_i-\bar{x})^2\right]=\sigma^2$$

Why Dividing by $n-1$? (Derivation)

$$\sum (x_i - \bar{x})^2 = \sum (x_i - \mu + \mu - \bar{x})^2$$
$$= \sum (x_i - \mu)^2 - n(\bar{x} - \mu)^2$$

Taking expectations:

$$E\left[\sum (x_i - \bar{x})^2\right] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Variability Measures: Standard Deviation

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

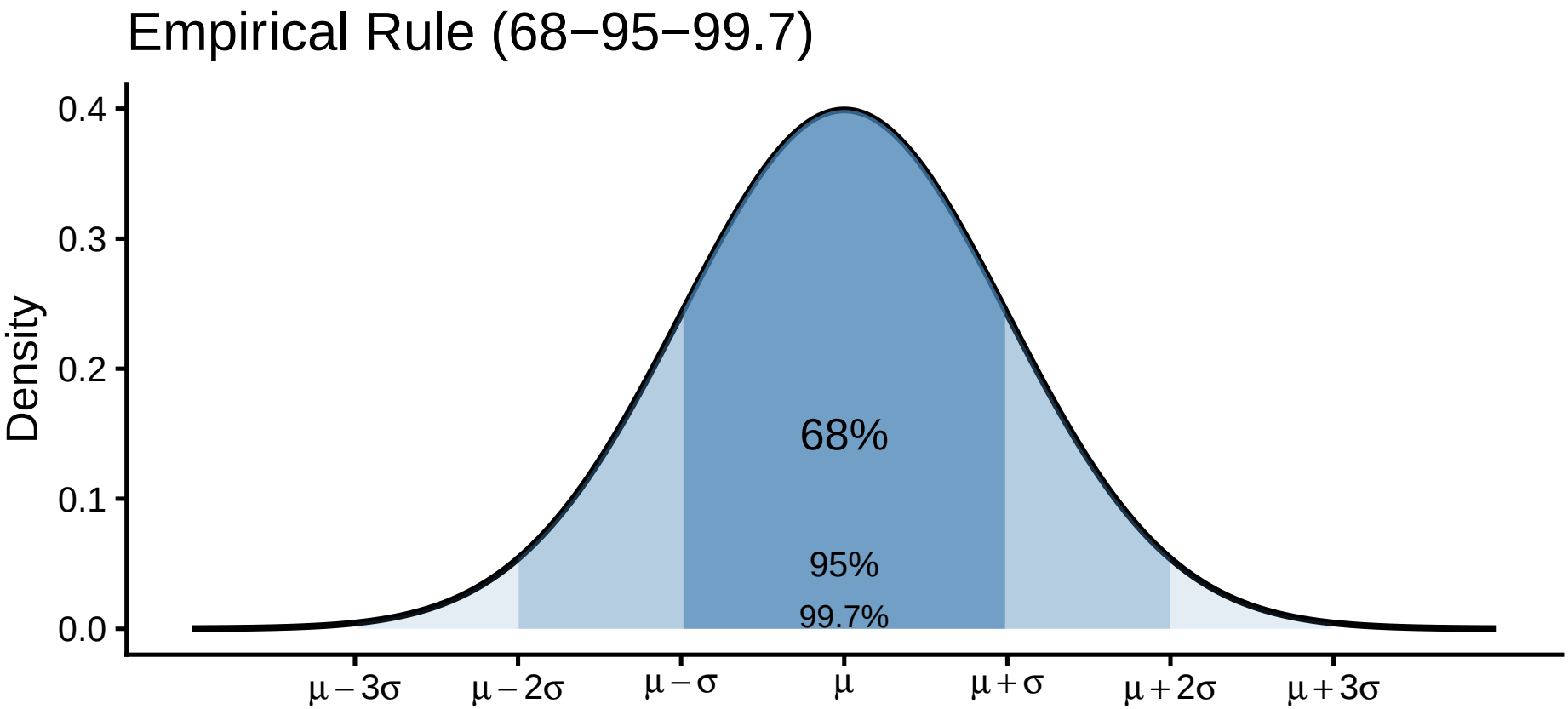
Variance is somewhat arbitrary — standard deviation has the **same units as the data**.

Chebyshev's Theorem

Regardless of how the data are distributed, at least $\left(1 - \frac{1}{k^2}\right)$ of values must fall within k standard deviations from the mean.

k	Minimum % within $k\sigma$
2	75%
3	89%

If the Distribution is Bell-Shaped...



Variability Measures: Standard Error

The **standard error** is the standard deviation of the sampling distribution of a statistic (usually the mean):

$$SE(\bar{x}) = \frac{\hat{\sigma}}{\sqrt{n}}$$

As sample size n increases, SE decreases — estimates become more precise.

Variability Measures: Standard Error — Visualized

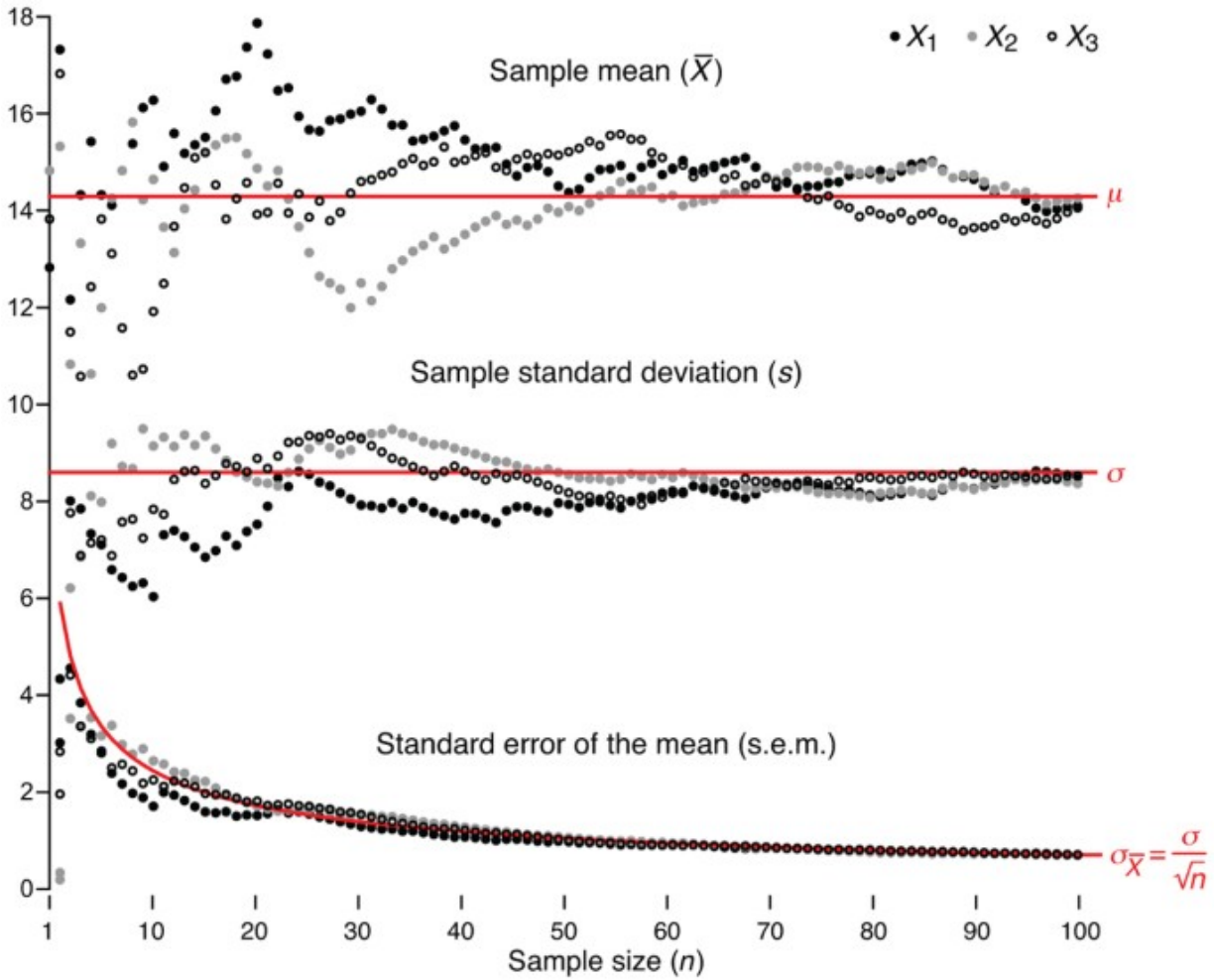


Figure 1: Sample mean, SD, and SE converge with increasing n

Variability Measures: Range

$$\text{Range} = x_{\max} - x_{\min}$$

Simple but sensitive to outliers.

Section 3

Relative Standing and Visualization

Relative Standing: Percentiles (Quantiles)

The nth percentile is a value such that $n\%$ of the observations fall at or below it.

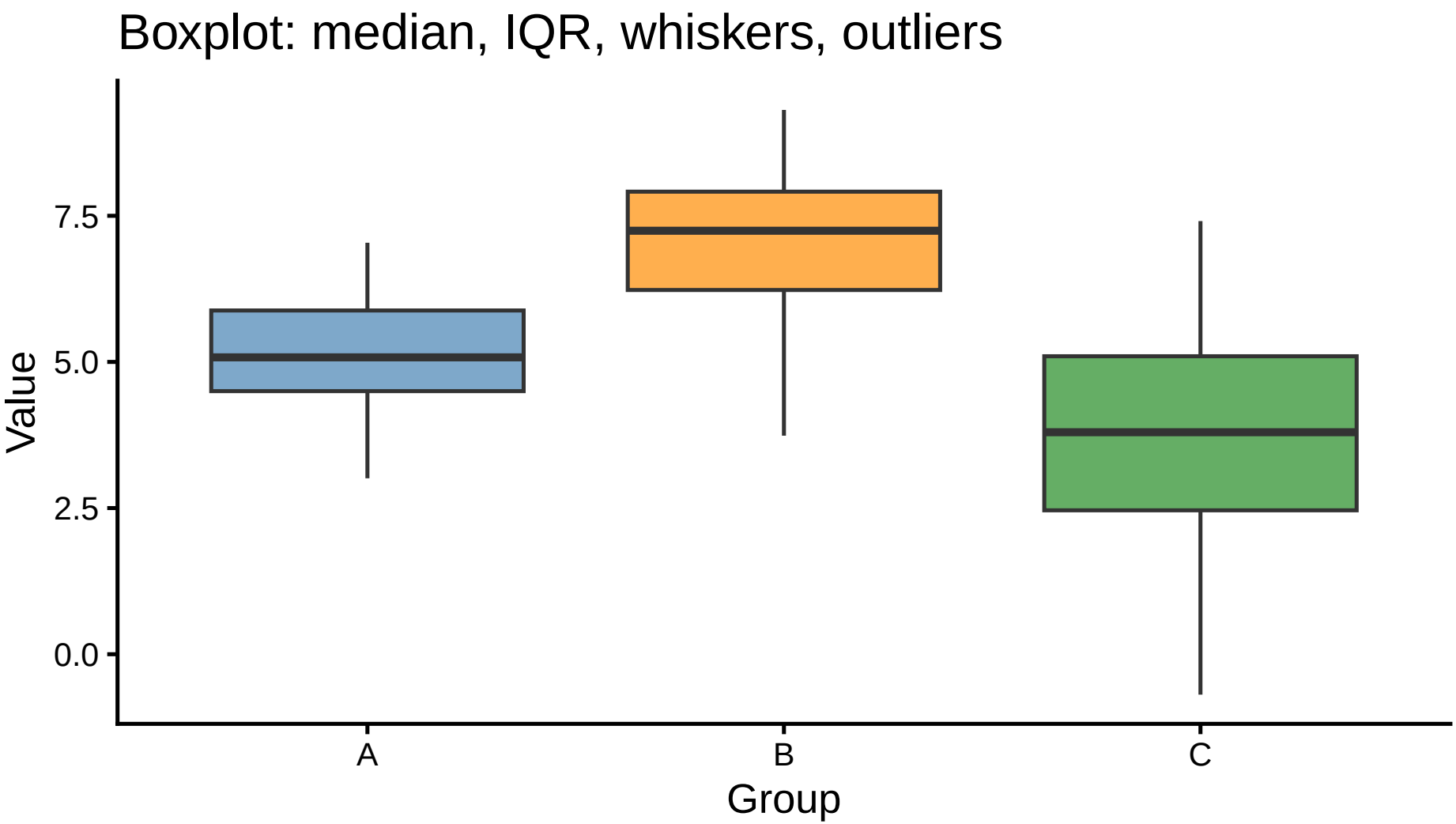
- $Q_1 =$ 25th percentile
- $Q_2 =$ 50th percentile $=$ **Median**
- $Q_3 =$ 75th percentile

Relative Standing: Quartiles and IQR

$$\text{IQR} = Q_3 - Q_1$$

Interquartile range — the range of the middle 50% of the data.

Summarizing Data — Boxplots



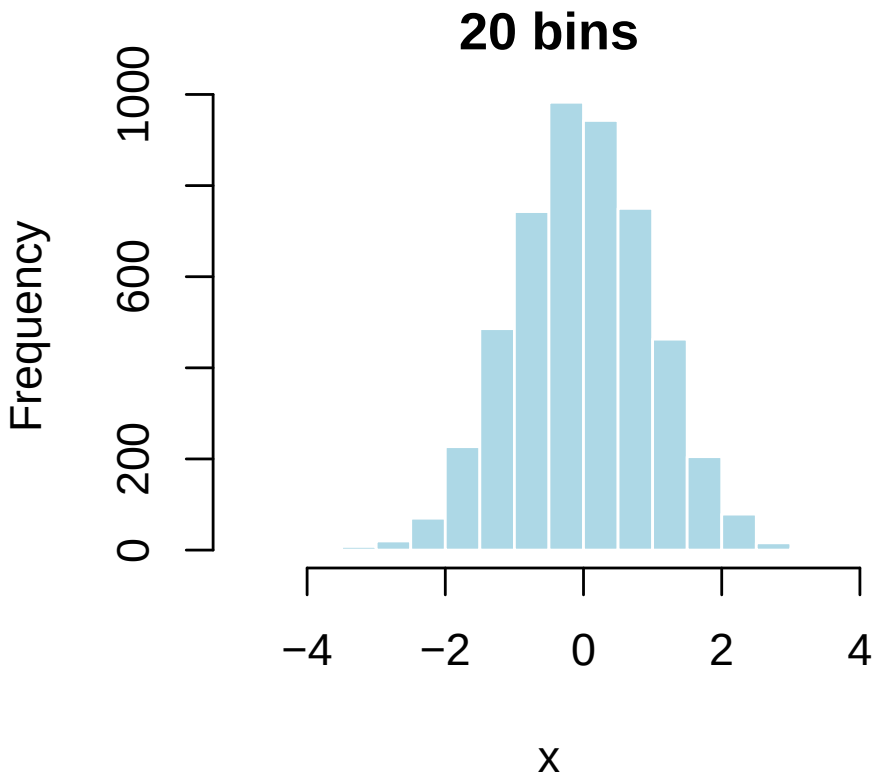
Summarizing Data — Boxplots

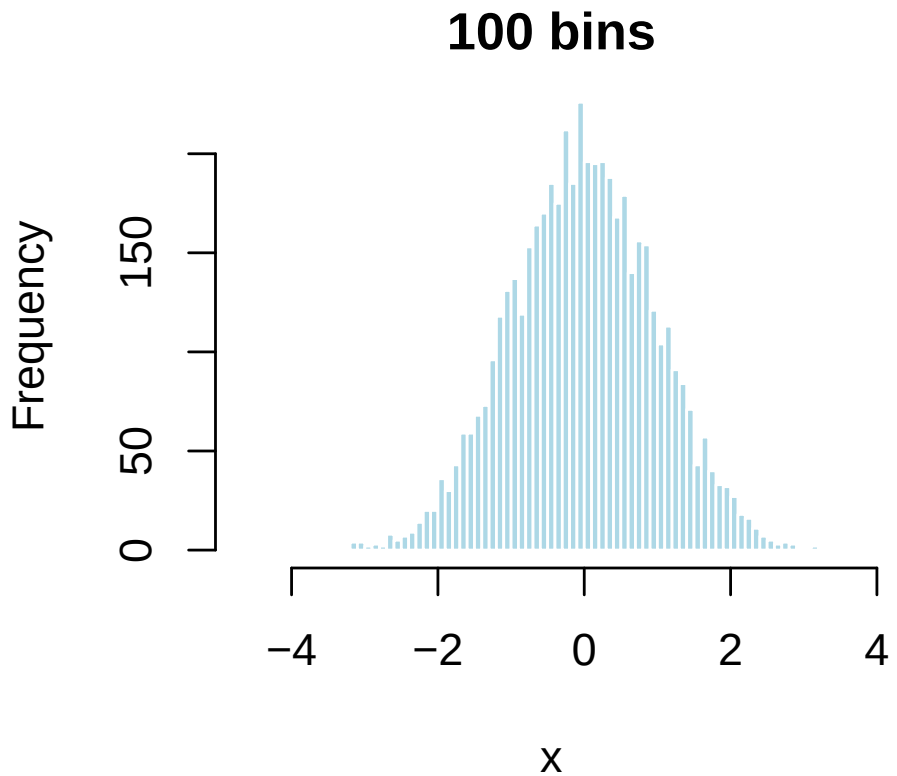
A boxplot shows:

- **Box:** Q_1 to Q_3 (IQR)
- **Line in box:** Median
- **Whiskers:** extend to $1.5\times$ IQR from the box
- **Points beyond whiskers:** potential outliers

Summarizing Data — Histograms

Used to visualize distribution (shape, center, range, variation) of continuous variables.





Section 4

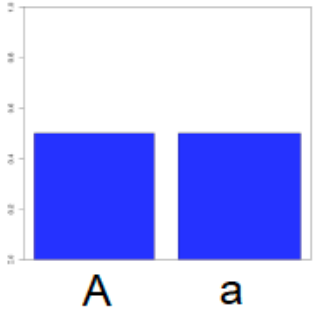
Probability Distributions

Probability Distributions of Random Variables

Discrete

Let X be a discrete rv. Then the *probability mass function (pmf)*, f(x), of X is:

$$f(x) = \begin{cases} P(X = x), & x \in \Omega \\ 0, & x \notin \Omega \end{cases}$$



Continuous

Let X be a continuous rv. Then the *probability density function (pdf)* of X is a function f(x) such that for any two numbers a and b with $a \leq b$:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

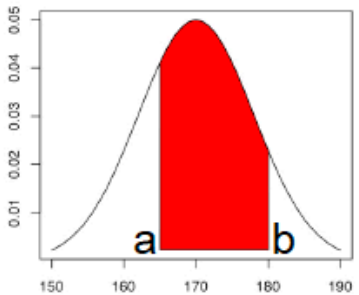
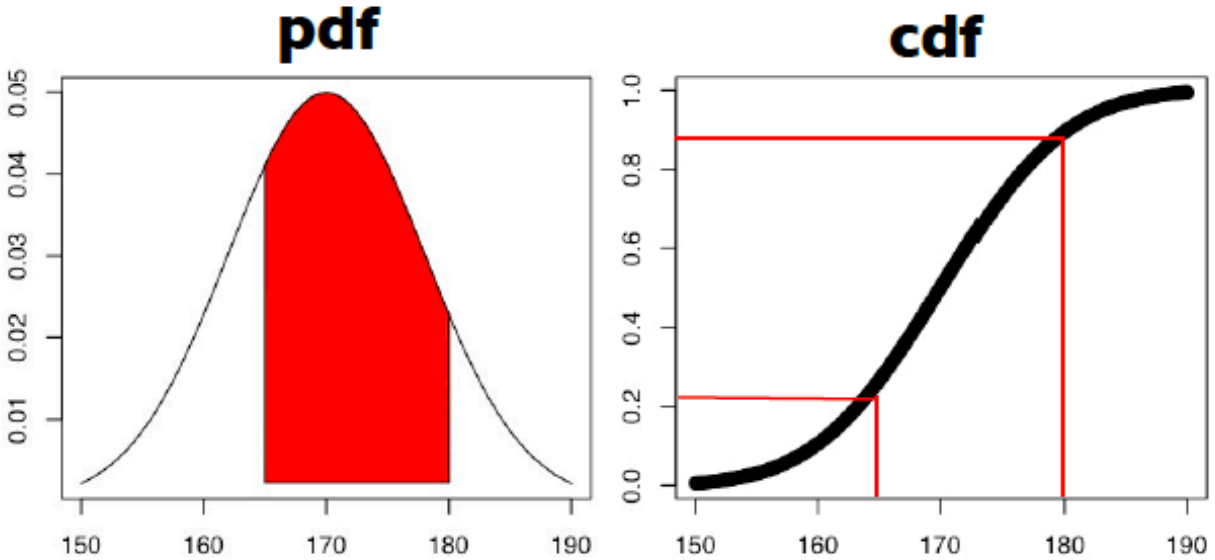


Figure 2: PMF (discrete) and PDF (continuous)

Using CDFs to Compute Probabilities

Continuous rv: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$



$$P(a \leq X \leq b) = F(b) - F(a)$$

Figure 3: PDF and CDF relationship

Expectation of Random Variables

$$E[X] = \begin{cases} \sum_x x \cdot P(X = x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} x \cdot f(x)\, dx & \text{(continuous)} \end{cases}$$

Variance of Random Variables

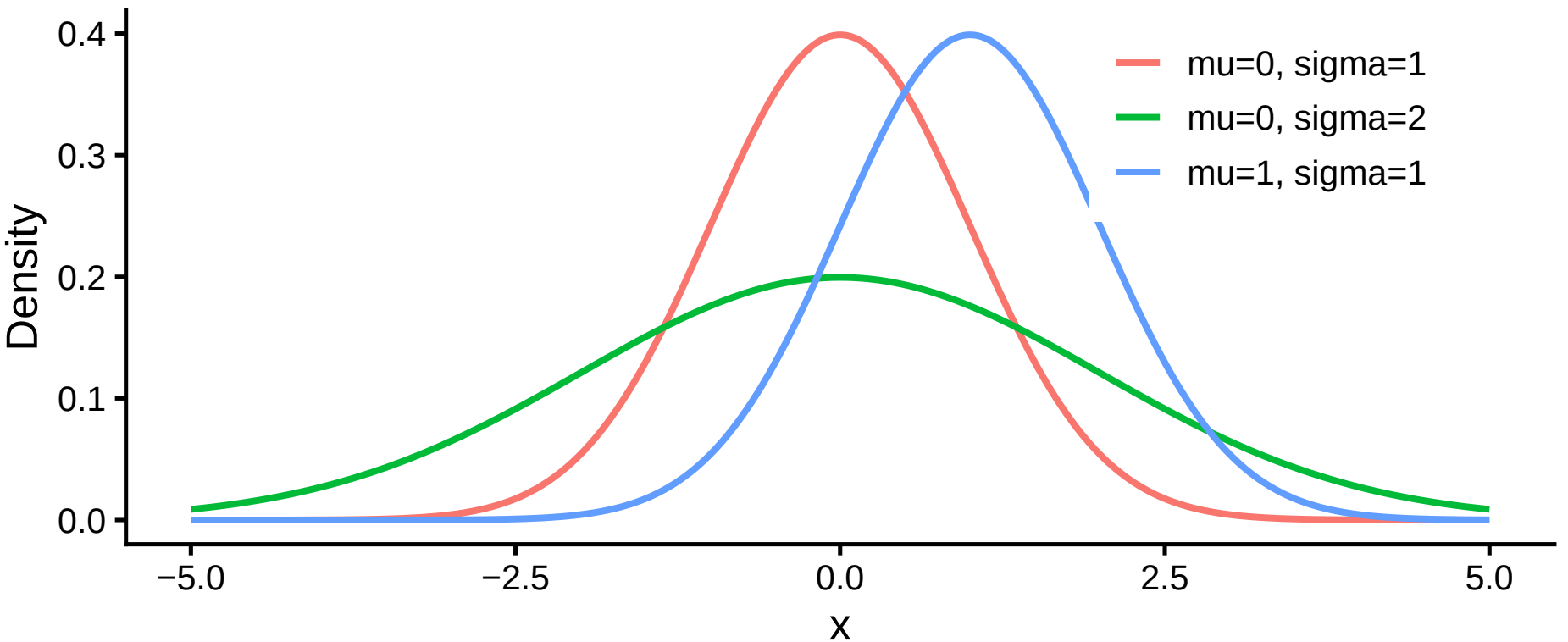
$$\mathsf{Var}(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$$

Section 5

Normal Distribution

Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Normal Distribution — Properties

- Symmetric around μ
- μ = mean = median = mode
- Defined by two parameters: μ (location) and σ (spread)
- Area under the curve = 1

Standardizing Normal RV

If $X \sim N(\mu, \sigma^2)$, we can standardize:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

This is the **standard normal distribution**.

Practical Utility of Normal Distribution

Many natural phenomena are approximately normal:

- Heights, weights
- Measurement errors
- Gene expression levels (often after log-transformation)

This makes the normal distribution the foundation for many statistical tests.

The Amazing Central Limit Theorem

CLT: Regardless of the population distribution, the sampling distribution of $\bar{X}$ approaches a normal distribution as $n \to \infty$:

$$\bar{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

The Central Limit Theorem — Visualized

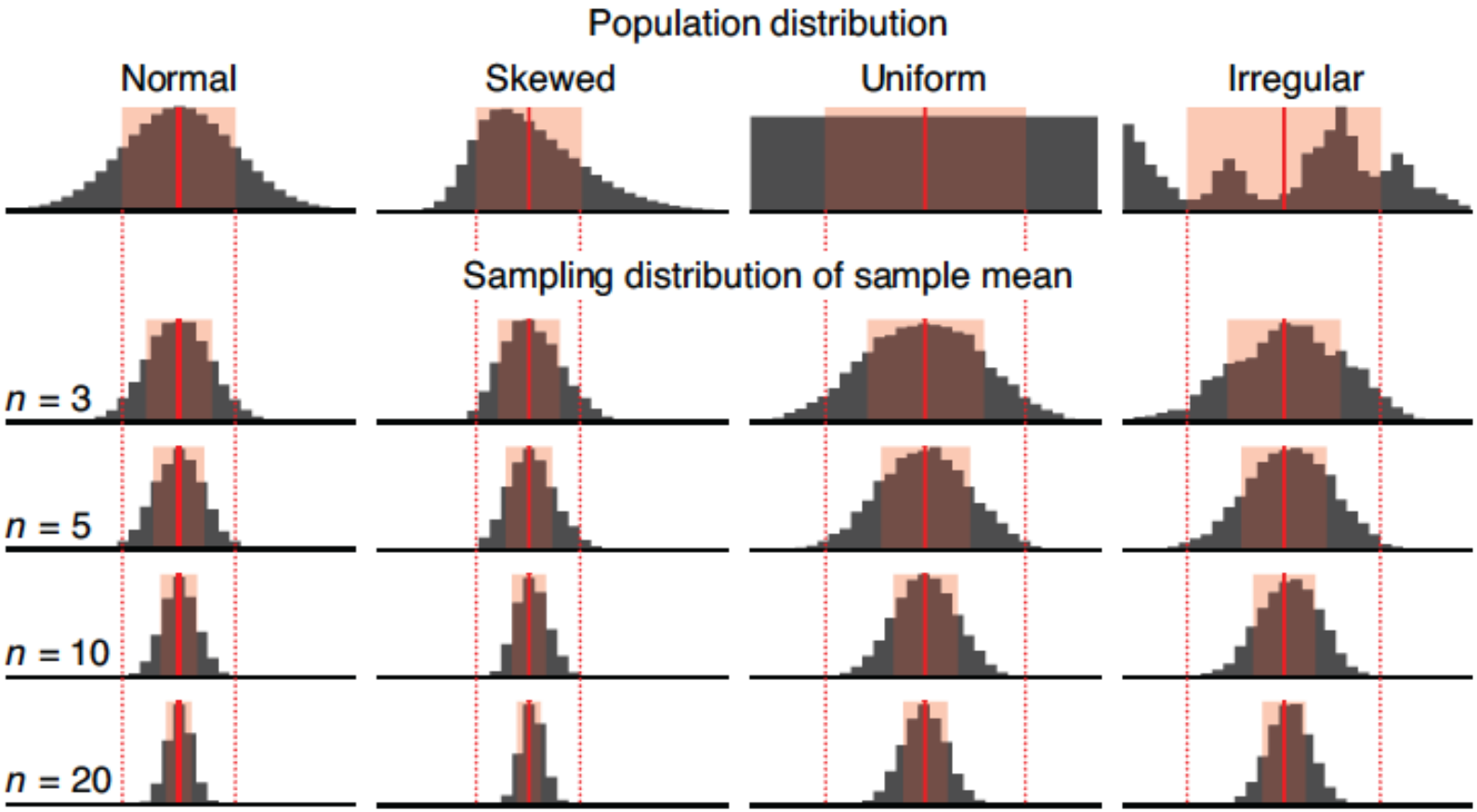


Figure 4: CLT: sampling distributions converge to normal regardless of population shape